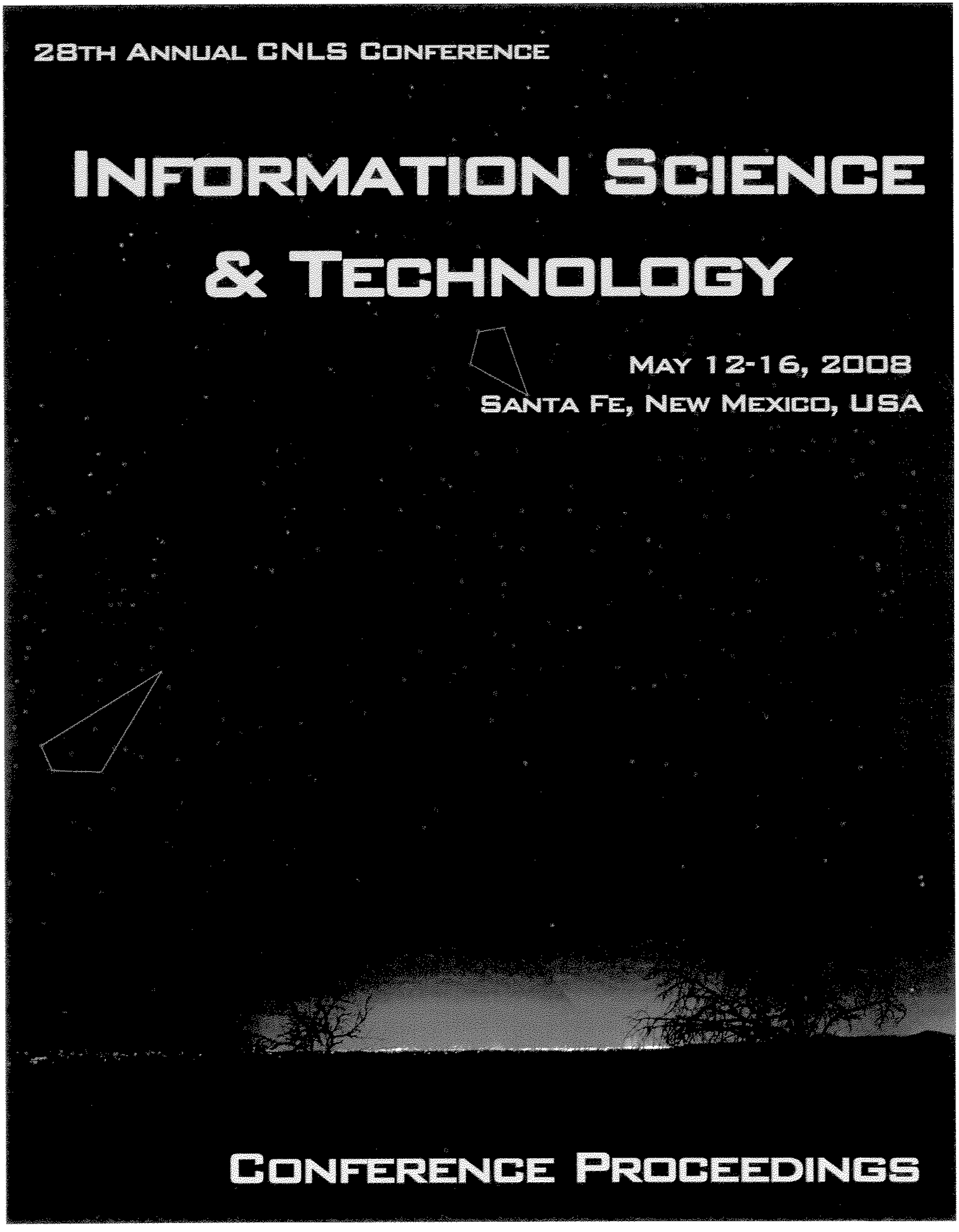
28th Annual CNLS Conference



Information Science & Technology

May 12-16, 2008
Santa Fe, New Mexico, USA

Conference Proceedings

INFORMATION SCIENCE AND TECHNOLOGY

May 12-16, 2008

Conference Schedule

Monday, May 12, 2008

Opening
9:00 AM – 9:15 AM

Conference Talks
9:15 AM – 4:20 PM

Tuesday, May 13, 2008

Conference Talks
9:00 AM – 4:20 PM

Reception/Cash Bar
6:00 PM – 6:30 PM

Dinner Banquet
6:30 PM – 8:30 PM

Wednesday, May 14, 2008

Conference Talks
9:00 AM – 12:40 PM

Thursday, May 15, 2008

Conference Talks
9:00 AM – 4:40 PM

Poster Reception
5:00 PM – 6:00 PM

Friday, May 16, 2008

Conference Talks
9:00 AM – 11:50 AM

Conference Organizers:
Frank Alexander, Dimitri Bertsekas, Luis Bettencourt, Ilya Nemenman, Pieter Swart, James Theiler, Mike Warren, Joanne Wendelberger

Conference Booklet Produced By: Kelle Ramsey, Adam Shipman, and Ellie Vigil

Center for
Nonlinear Studies

***This Event is
Open To the Public***

Los Alamos
NATIONAL LABORATORY
EST.1943

Monday, May 12, 2008

9:00 am - 9:15 am	**Opening Remarks** – Bob Ecke, Center for Nonlinear Studies, Director
9:15 am - 9:30 am	**Information Science and Technology at LANL** – Terry Wallace, Principal Associate Director for Science, Technology and Engineering
9:30 am - 9:40 am	**Conference Remarks** – Frank Alexander, Information Science and Technology, Acting Center Leader
9:40 am - 9:50 am	**Conference Details** – Adam Shipman, Center for Nonlinear Studies, Conference Coordinator
9:50 am - 10:40 am	**Grace Wahba** (University of Wisconsin) *The LASSO-Patternsearch Algorithm: Finding "Patterns in a Haystack"*
10:40 am - 11:00 am	**Break**
11:00 am - 11:50 am	**Alan Willsky** (Massachusetts Institute of Technology) *Problems and Challenges that Keep Me up at Night (in a Good Way)*
11:50 am - 1:30 pm	**Lunch Break**
1:30 pm - 2:20 pm	**Greg Taylor** (University of New Mexico) *The Long Wavelength Array*
2:20 pm - 3:10 pm	**Allan Wilks** (AT&T) *Waterworks for Stream Processing*
3:10 pm - 3:30 pm	**Break**
3:30 pm - 4:20 pm	**Stephen Boyd** (Stanford University) *Convex Optimization in Large-Scale and Real-Time Data-Driven Applications*

The LASSO-Patternsearch Algorithm: Finding "Patterns in a Haystack"

Grace Wahba
(University of Wisconsin-Madison)

The LASSO-Patternsearch Algorithm: Finding "patterns in a haystack." We describe the LASSO-Patternsearch algorithm, whose core is the application of a global LASSO (l_sub 1 penalized likelihood) to Bernoulli response data given a very large attribute vector from a restricted multivariate Bernoulli distribution which is unknown but assumed to be sparse. That is, the log linear expansion of this (unknown) distribution is assumed to have very few terms, some of which, however may be of higher order (patterns). Tuning methods for Bernoulli data are scarce in the literature for good reason.

We use a modification (BGACV) of an approximately unbiassed risk estimator (GACV) for Bernoulli data to tune the LASSO, and a final fitting step. An algorithm, which can handle a very large number of candidate patterns in a global optimization scheme is given. Applications to demographic and genetic data are described.

Problems and Challenges that Keep Me up at Night (in a Good Way)

Alan Willsky
(MIT - LIDS/EECS)

This presentation begins by presenting four major challenges to large-scale data analysis and information extraction that represent the current drivers behind our research. These include large-scale geophysical data assimilation and analysis; dynamic tracking of multiple entities and the discovery of complex links among them; learning from and assisting human experts in extracting information and navigating through complex data sets; and decision-making to support sensing and inference in distributed and complex environments. We describe some of the characteristics of these "closet monsters" that we feel are important and that should provide drivers for research for quite some time into the future. In the remainder of this talk we describe some of our group's lines of research aimed at taking steps toward developing the understanding and additions to our theoretical and methodological "toolkits" required to confront challenges such as these. Topics on which we will touch include several new methods for estimation and optimization in large-scale graphical models, machine learning for extracting links among entities and for learning salient aspects of statistical dependencies in high-dimensional data when the objectives are far simpler (e.g., hypothesis testing), and extracting geometry and dynamically evolving geometry in large-scale random fields. We close with a few thoughts on directions we're pursuing as we inch toward our closet door and confront those monsters head-on.

The Long Wavelength Array

Greg Taylor
(University of New Mexico)

The Long Wavelength Array (LWA; http://lwa.unm.edu) will be a powerful, low frequency telescope emphasizing high angular resolution exploration of the Universe. It is almost entirely through advances in information science that this region of the spectrum can be opened up for astrophysical study. The LWA will require stations distributed over long baselines (~400 km), thousands of elements, terabytes of data processed each second, and significant advances in our understanding of the ionosphere and low frequency imaging. The result will be a versatile, user-oriented electronic array with excellent sensitivity and angular resolution. I will review the current status of the project and highlight some of the challenges.

Waterworks for Stream Processing

Allan Wilks
(AT&T Labs Research)

Over the years, Rick Becker and I have constructed a number of generic tools for dealing with large streams of data at AT&T. I will talk about several useful paradigms we have worked with (some water-related!) that have helped us to build robust, lightweight, flexible software, much of which has been deployed in some of the company's most critical applications, such as fraud detection and billing.

Convex Optimization in Large-Scale and Real-Time Data-Driven Applications

Stephen Boyd
(Stanford University)

Convex optimization is now widely used in control, signal processing, networking, communications, machine learning, finance, combinatorial optimization, and other fields. For many problem classes reliable general purpose solvers are now available, with development of new algorithms and implementations continuing at a rapid pace. In this talk I will given an overview of some recent advances. The first is the development of specification and modeling languages specifically for convex optimization. These languages allow very rapid development of applications based on convex optimization, and enhance learning and teaching of the methods. The second is the development of methods for large convex problems, with millions (or more) of variables and constraints, for specific families of problems arising in applications. Truncated Newton interior-point methods, with well-chosen pre-conditioner, can solve far larger problems than generic methods. The third advance is in the area of algorithms for fast solution of convex optimization problems, for use in real-time and embedded applications.

TUESDAY, MAY 13, 2008

9:00 am - 9:50 am	**Katy Borner** (Indiana University) *Science Maps in Action*
9:50 am - 10:40 am	**Doug Nychka** (National Center for Atmospheric Research) *Challenges of Regional Climate Modeling and Validation*
10:40 am - 11:00 am	**Break**
11:00 am - 11:50 am	**Ingo Steinwart** (Los Alamos National Laboratory) *Machine Learning with Kernel Methods: Some Recent Results and Open Questions*
11:50 am - 1:30 pm	**Lunch**
1:30 pm - 2:20 pm	**Sergei Pond** (University of California, San Diego) *Evolutionary Fingerprinting of Viral Genes*
2:20 pm - 3:10 pm	**Stefano Monti** (Broad Institute) *Cancer Genomics: From Integrative Analysis to Targeted Therapy*
3:10 pm - 3:30 pm	**Break**
3:30 pm - 4:20 pm	**Vijay Nair** (University of Michigan) *Statistical Inverse Problems in Network Tomography*
6:00 pm – 6:30 pm	**Reception/Cash Bar**
6:30 pm – 8:30 pm	**Banquet**

Science Maps in Action

Katy Borner
(Indiana University)

Cartographic maps of physical places have guided mankind's explorations for centuries. They enabled the discovery of new worlds while also marking territories inhabited by unknown monsters. Domain maps of abstract semantic spaces, see http://scimaps.org, aim to serve today's explorers' understanding and navigating the world of science. The maps are generated through scientific analysis of large-scale scholarly datasets in an effort to connect and make sense of the bits and pieces of knowledge they contain. They can be used to objectively identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. Local maps provide overviews of a specific area: its homogeneity, import-export factors, and relative speed. They allow one to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research. Global maps show the overall structure and evolution of our collective scholarly knowledge. This talk will present an overview of the techniques and cybertechnologies used to study science by scientific means together with sample science maps and their interpretations.

Challenges of Regional Climate Modeling and Validation

Douglas Nychka
(National Center for Atmospheric Research)

As attention shifts from broad global summaries of climate change to more specific regional results there is a need for statistics to analyze observations and model output that have significant variability and also to quantify the uncertainty in regional projections. This talk will survey some work on interpreting regional climate experiments. In large multi-model studies one challenge is to understand the contributions of different global and regional model combinations to the simulated climate. This is difficult because the individual runs tend to be short in length. Thus one is faced with the paradox of generating massive data sets that still demand statistical analysis to quantify significant features. We suggest some spatial models for the climate fields based on sparse approximations to the covariance matrix and derive an ANOVA like decomposition for the fields. The decomposition into main effects and interactions helps to isolate the effects of different models. The spatial models provide a rigorous framework for assessing statistical significance and comparing simulations to observed climate. This approach is illustrated for output from the PRUDENCE program and we also discuss the newer NARCCAP experiments for the regional climate of North America.

Machine Learning with Kernel Methods: Some Recent Results and Open Questions

Ingo Steinwart
(Los Alamos National Laboratory)

Kernel methods, such as support vector machines, represent highly successful machine learning algorithms. The first part of this talk presents some recent results on these methods. In particular, the role of the loss function, the structure of the output function, and the generalization performance are reviewed. The second part of this talk then discusses some open questions, such as the empirically observed but theoretically not understood adaptivity to high dimensional input sources, and the handling of non-i.i.d. data.

Evolutionary Fingerprinting of Viral Genes

Sergei Pond
(University of California, San Diego)

Over time, natural selection molds every gene into a unique mosaic of sites evolving rapidly and resisting change – an 'evolutionary fingerprint' of the gene. We introduce a metric, called the evolutionary selection distance (ESD), to identify similarities and idiosyncrasies in selection pressure between sequence alignments. Using a broad survey of viral genes, we apply a variety of computational techniques employing ESD to classify genes by the similarity of their evolutionary fingerprints, identify genes with distinctive evolutionary features, and correlate evolution with phylogenetic, functional, and taxonomic information.

We demonstrate that genes within the same functional group tend to exhibit similar evolutionary patterns, both within a single viral genome and between different viruses; similarity in selection pressures mirrors phylogenetic relationships among hepatitis C virus subtypes, but not HIV-1 subtypes and that evolutionary patterns in the hemagglutinin gene of the Influenza A virus are largely determined by the host with a few notable exceptions. By comparing genes at the level of the evolutionary processes rather than the pattern of sequence variation, we can compare both closely and distantly related genes, potentially revealing the guiding principles underlying the evolution of genetic diversity.

Cancer Genomics: From Integrative Analysis to Targeted Therapy

Stefano Monti
(Broad Institute)

One of the striking features of the "genomics" era is that a significant part of biological research has moved from the wet lab to the computer lab. The shift is driven by the introduction of high-throughput technologies, such as genome-wide expression profiling and high-density SNP arrays, which allow for the simultaneous measurement of tens (hundreds) of thousands of biological variables, whose integration and interpretation present considerable computational challenges.

In this talk, I will outline some of the outstanding challenges, describe some of the data types and the computational techniques used for their processing and analysis, and show how they can be used toward the advancement of cancer study and targeted therapy.

Statistical Inverse Problems in Network Tomography

Vijay Nair
(University of Michigan, Ann Arbor)

The term network tomography deals with several large-scale inverse problems that arise in the modeling and analysis of computer and communications networks. They include: a) the estimation of origin-destination (end-to-end) traffic matrix from data collected at the individual nodes, b) recovering link-level parameters, such as packet loss rates and delay distributions, from end-to-end path-level measurements; and c) identifying the network topology from end-to-end path-level data. This talk will provide an overview of these problems and describe applications to analyzing and monitoring computer and communications networks. This is joint work with Earl Lawrence, George Michailidis, and Xiaodong Yang.

Wednesday, May 14, 2008

9:00 am - 9:50 am	**Diane Lambert** (Google) *Statistical Analysis at Google Speed and Scale*
9:50 am - 10:40 am	**Alex Smola** (Australia National University) *Painless Distribution Representations*
10:40 am - 11:00 am	**Break**
11:00 am - 11:50 am	**Lawrence Saul** (University of California, San Diego) *Statistics, Geometry, Computation: Searching for Low Dimensional Structure in High Dimensional Data*
11:50 am - 12:40 pm	**Misha Chertkov** (Los Alamos National Laboratory) *Loop Calculus for Graphical Models*
	Day Ends Early

Statistical Analysis at Google Speed and Scale

Diane Lambert
(Google)

Google is an enormous network of data. This talk will describe some of the statistical challenges and opportunities involved in routinely turning that data into useable information.

Painless Distribution Representations

Alexander Smola
(NICTA and Australian National University)

Entropy and Mutual Information are popular tools for modeling data. In this talk I present alternative methods for comparing distributions which do not require density estimation but which rely on convergence properties of the expectation operator instead.

This allows us to design algorithms for two-sample tests, independent component analysis, density estimation, clustering, feature selection, low-dimensional data representation and related unsupervised problems based on a unifying framework. It turns out that a large number of existing algorithms, ranging from sorting, k-means clustering, (kernel) principal component analysis, maximum variance unfolding, Pearson correlation to the Kolmogorov-Smirnov test and the earth mover's distance are special cases of our framework.

Statistics, Geometry, Computation: Searching for Low Dimensional Structure in High Dimensional Data

Lawrence Saul
(University of California, San Diego)

How can we detect low dimensional structure in high dimensional data? If the data is mainly confined to a low dimensional subspace, then simple linear methods can be used to discover the subspace and estimate its dimensionality. More generally, though, if the data lies on (or near) a low dimensional manifold, then its structure may be highly nonlinear, and linear methods are bound to fail.

In this talk, building on elementary ideas from convex optimization, spectral graph theory, and differential geometry, I will describe an algorithm that we have recently developed for this problem. Given high dimensional data sampled from a low dimensional manifold, our algorithm can be used to estimate the data's intrinsic dimensionality and to compute a faithful low dimensional representation. Surprisingly, the main computations in our approach are based on highly tractable optimizations, such as nearest-neighbor searches, least squares fits, eigenvalue problems, and semidefinite programming. In practice, the results from our algorithm are quite useful for the visualization and analysis of high dimensional data sets. I will discuss several applications of our work, as well as open questions for future research.

Loop Calculus for Graphical Models

Misha Chertkov
(Los Alamos National Laboratory)

Loop Calculus introduced in [Chertkov, Chernyak '06] constitutes a new theoretical tool that expresses explicitly the symbol Maximum-A-Posteriori solution of a general statistical inference problem via a solution of the Belief Propagation, or Bethe-Pieirls, (BP) equations. This finding brought a new significance to the BP concept, which in the past was thought of as just a loop-free approximation. In this presentation I will explain main concept and feature challenges of the Loop Calculus approach. I will also discuss algorithmic utility of the Loop Calculus for improved decoding of graphical codes, inference on planar graph and matching of particles/trajectories in chaotic flows.

Thursday, May 15, 2008

9:00 am - 9:50 am	**Bud Mishra** (New York University) *S*M*A*S*H: Single Molecule Approaches to Sequencing by Hybridization*
9:50 am - 10:40 am	**Naftali Tishby** (Hebrew University) *Predictive Information and the Perception-Action-Cycle*
10:40 am - 11:00 am	**Break**
11:00 am - 11:50 am	**Jose Principe** (University of Florida) *On-Line Kernel Learning*
11:50 am - 1:00 pm	**Lunch**
1:00 pm - 1:50 pm	**Alan Schaum** (Naval Research Laboratory) *Hyperspectral Image Processing: The Extra-Sensory Solution*
1:50 pm - 2:40 pm	**Alex Szalay** (Johns Hopkins University) *Peta-Scale Data-Intensive Computing*
2:40 pm - 3:00 pm	**Break**
3:00 pm - 3:50 pm	**Michael Jordan** (University of California, Berkeley) *Nonparametric Bayesian Graphical Models*
3:50 pm - 4:40 pm	**Benjamin Van Roy** (Stanford University) *Decentralization and Message-Passing*
5:00 pm - 6:00 pm	**Poster Session**

S*M*A*S*H: Single Molecule Approaches to Sequencing by Hybridization

Bud Mishra
(Courant Institute)

SMASH is a technology for sequencing a human size genome of 6 Gigabases (including both haplotypes) without using any prior sequence information. We have aimed the technology for eventually (e.g., in less than a decade) achieving a competitively low cost for each genome sequence produced (e.g. US$1000 or less), while assuring a high quality (e.g., standard of "high quality draft sequence" similar to the mouse genome sequence published in December 2002). This technology is hoped to play a significant disruptive role in the future predictive personalized biomedicine as well as other areas of biotech industries.

These goals require successful integration of three different component technologies: (1) Optical Mapping to create Ordered Restriction Maps with respect to an enzyme, (2) Hybridization of a pool of oligonucleotide probes (LNA probes) with Single Genomic dsDNAs, and (3) Algorithms to solve "localized versions" of PSBH (Positional Sequencing by Hybridization) problems over the whole genome.

Unlike many of its competitors, the technology works with small amount of genomic materials, operates top-down, employs a Bayesian algorithm to create haplotypic sequence assembly without an auxiliary shotgun assembler, tolerates noise in the data well and is cost-effective at multiple scales. By construction, it avoids errors due to hompolymeric runs, haplotypic ambiguities and large-scale rearrangement errors. Its scientific feasibility has been demonstrated through many important algorithmic, chemical, and mathematical innovations over the last two years, further reassuring the soundness of the principles, science, and strategy for technology development.

Predictive Information and the Perception-Action-Cycle

Naftali Tishby
(Hebrew University)

Biology is all about the ability of organisms to exploit the partial predictability of their environment. In this talk I will try to justify and quantify this statement, by drawing some rigorous analogies between Shannon's coding theorems and the cost-value tradeoff that organisms must deal with. In particular, I will argue that optimal adaptation to the environment can be equivalently formulated as a tradeoff between affordable predictive information, collected by the sensors in the past, and valuable information about the future, materialized by decisions, actions and rewards. This suggests a concrete computational paradigm for biological adaptation that can be experimentally tested and evaluated.

On-Line Kernel Learning

Jose C. Principe, Weifeng Liu
(University of Florida)

This talk describes how Kernel Least Mean Squares (KLMS) is well posed in the sense of Hadamard, and therefore does not need explicit regularization. In fact the stepsize or learning rate works as a regularizer. Alternatively, this result is a solid foundation for early stopping, which is very popular in neural network learning. The class of kernel affine projection algorithms (KAPA) shares the same property. As a consequence, much simpler on-line kernel algorithms are possible using this methodology. We are interested in nonlinear signal processing applications, but the results are applicable to machine learning problems too.

Hyperspectral Image Processing: The Extra-Sensory Solution

Alan Schaum
(Naval Research Laboratory)

Large-format cameras used for wide-area surveillance can capture orders of magnitude more data than can be examined in a timely way by a small team of analysts. One response to this problem examines details inside the light from each pixel, to extract content not accessible to human senses. This can require the collection of even larger data volumes, but the enriched data can be analyzed with advanced algorithms that automatically flag interesting pixels. Good detection methods can more than offset the associated increase in data rates. However, overall false alarm rates often remain too high for fully autonomous operation.

Imaging spectrometers, for example, measure a full spectrum at each pixel, collecting hundreds of times more data than panchromatic cameras. When the spectrometers are integrated with autonomous techniques for detection, they become hyperspectral imaging (HSI) systems, which may one day make wide area panchromatic sensing obsolete. Here we describe the informational challenges facing this technology. Most of the proven HIS detection algorithms can be understood as a comparison of hypothesized information content in each pixel. We describe some advanced approaches that push this principle to the limit, using an intuitive geometrical interpretation in a high-dimensional spectral space.

Peta-Scale Data-Intensive Computing

Alex Szalay
(Johns Hopkins University)

Scientific data is doubling every year. Virtual Observatories are established over every scale of the physical world: from elementary particles to materials, biological systems, environmental observatories, remote sensing, and the universe. These collaborations collect increasing amounts of data, often close to a rate of petabytes per year. Many scientists will soon obtain most of their data from large scientific repositories of data, often stored in the form of databases. The talk will discuss the different requirements for such databases, and discuss user behavior in a few concrete examples taken from astronomy, in particular from the 6 year usage of the Sloan Digital Sky Survey database. Interesting query patterns are emerging, where users create custom “crawlers” to break large queries into many repetitive ones. The trial-and-error behavior of many exploratory projects will be also discussed. The talk will also present various scalable alternatives to large scientific analysis facilities.

Nonparametric Bayesian Graphical Models

Michael Jordan
(University of California, Berkeley)

Graphical models have classically been developed within the realm of parametric statistics---the distributions that have been considered are the multinomial, the Gaussian and other distributions in the exponential family, and the set of graphs under consideration in a given problem is generally taken to be fixed and finite. In this talk I overview some of the stochastic processes that are allowing us to move beyond these classical parametric restrictions. These include the Dirichlet process, the beta process, the gamma process, and various recursive constructions (e.g., Bayesian hierarchies) involving these processes. I discuss some of the marginals that arise from integrating out these stochastic processes. These are guaranteed to enhance your appetite for nonparametrics. I also present applications of models based on these stochastic processes to problems in structural biology, statistical genetics, natural language processing and computational vision.

[Joint work with Yee Whye Teh and Romain Thibaux.]

Decentralization and Message-Passing

Benjamin Van Roy
(Stanford University)

In many complex systems there is so much data from so many sources driving so many decisions that computations must be decentralized. In principle, an individual component can make system-optimal local decisions given an externality function, which captures the impact of its decisions on the system objective. It is generally too complex to compute or even encode exact externality functions, and as such, one might instead settle for shadow prices (i.e., Lagrange multipliers), which provide linear approximations. We propose the use of general separable approximations, which can capture nonlinear effects while escaping the intractability of exact externality functions. We discuss how the min-sum message passing algorithm can be used to compute such approximations. As an example, we design a distributed protocol for rate control in a communication network with inelastic traffic.

Friday, May 16, 2008

9:00 am - 9:50 am	**Paul Ginsparg** (Cornell University) *Next-Generation Implication of Open Access*
9:50 am - 10:40 am	**Jennie Si** (Arizona State University) *Cortical Neural Coding for Decision Making and Control Strategy Development*
10:40 am - 11:00 am	**Break**
11:00 am - 11:50 am	**Peter Quinn** (University of Western Australia) *The Square Kilometre Array - Too Much Data?*

Next-Generation Implication of Open Access

Paul Ginsparg
(Cornell University)

True open access to scientific publications not only gives readers the possibility to read articles without paying subscription, but also makes the material available for automated ingestion and harvesting by 3rd parties. Once articles and associated data become universally treatable as computable objects, openly available to 3rd party aggregators and value-added services, what new services can we expect, and how will they change the way that researchers interact with their scholarly communications infrastructure? I will discuss straightforward applications of existing ideas and services, including full text mining, citation analysis, collaborative filtering, external database linkages, interoperability, and other forms of automated markup, and speculate on the sociology of the next generation of users. Many of these applications involve working with datasets only formerly considered massive in scale, but which still require algorithmic heuristics.

Cortical Neural Coding for Decision Making and Control Strategy Development

Jennie Si
(Arizona State University)

This talk addresses the question and elaborates the challenges of how a rat develops its control strategy in a behavioral task by making correct decisions to successfully complete his trials. An apparatus is used to record a rat's behavioral and neural activities while he performs a directional control task. The experiment involves a self-paced, freely moving rat trying to switch a series of light positions to a center location by pressing a left or rtemp1234
ight control lever. The rat works for a food reward. I hypothesize that the rat's control strategy of optimizing his chance of food reward is represented in the rat's motor cortical neural spike activities. But how the patterns of neural activities distributed across populations of cells encode important information as abstract as a control strategy? Where and how the control strategy information is represented in the neural activities? I'll discuss approximate implementations of dynamic programming (DP), a computational approach to finding an optimal policy by employing the principle of optimality, as possible means of uncovering the neural basis of decision making and control strategy development in behaving animals. Properties of the approximate DP models will be discussed, as well as feasibilities of these models to interpret the connection between cortical neural activities and animal decision and control behaviors.

The Square Kilometre Array - Too Much Data?

Peter Quinn
(University of Western Australia)

The Square Kilometre Array (SKA) is a global science project which has 50 times the collecting area, 100 times the field of view and 2 times the bandwidth of any existing radio telescope. That represents an increase in "discovery space" by a factor of 10,000. This kind of increase in scientific capabilities far exceeds the previous greatest step forward when Galileo first used a telescope to look at sky 400 years ago. The data processing challenges, data complexity and data volume presented by the SKA will require hyper-Moore increases in CPU and storage technologies. Even allowing for compliant hardware, the software and data storage complexity will require new approaches to astronomical data exploration. I will review the SKA project and its projected data and processing needs in the light of current key science questions and the experience we will have to draw upon with current survey telescopes.